Some Back-of-the-envelope Calculations for Surveillance Testing

Christina, Frank, and Varun

11/09/2020

About This Document

This document outlines some simple strategies to assess the prevalence of coronavirus when there is insufficient capacity to test everyone in the population of interest. The ideas we discuss are neither novel nor original, and this document does not propose a fully-detailed plan for surveillance testing. Our goals are merely to explain the intuition behind randomized testing and pooled testing, and give a rough approximation to the number of tests required to provide useful information. To keep the discussion as simple as possible, we begin by abstracting from the problem of false negatives and false positives. The final section discusses how these affect our analysis.

The Bottom Line

Say something like x number of tests per week could reasonably work if prevalence is such-and-such (with and without pooling)

Random Testing

As a rough approximation, suppose that there are $N = 20{,}000$ undergraduate and taught masters students in residence at Oxford during MT. An unknown number C of them have coronavirus. With a perfect test, we could learn C exactly by conducting 20,000 tests. But what if we only have the capacity to carry out n tests, where n is much smaller than 20,000? In this case we cannot learn C exactly, but by employing *random testing* we can produce an estimate that is sufficiently accurate to guide our mitigation policies.

The most basic form of random testing relies on a *simple random sample*. To use this procedure, we place all 20,000 student id numbers into a hat, mix them up thoroughly, and then draw out n at random. These n students are sent for a coronavirus test. We then use the number of students in our *sample* who have coronavirus to estimate the number of students in the *population* who have coronavirus. For example, if one student in a sample of 2000 tests positive, then we would estimate that there are there are 10 cases in the University as a whole. Expressed as a formula, if P is the number of students in our sample who test positive, then our estimate E of the total number of cases in the university equals P/n multiplied by N. In the example, 10/2000 multiplied by 20,000 equals 10.

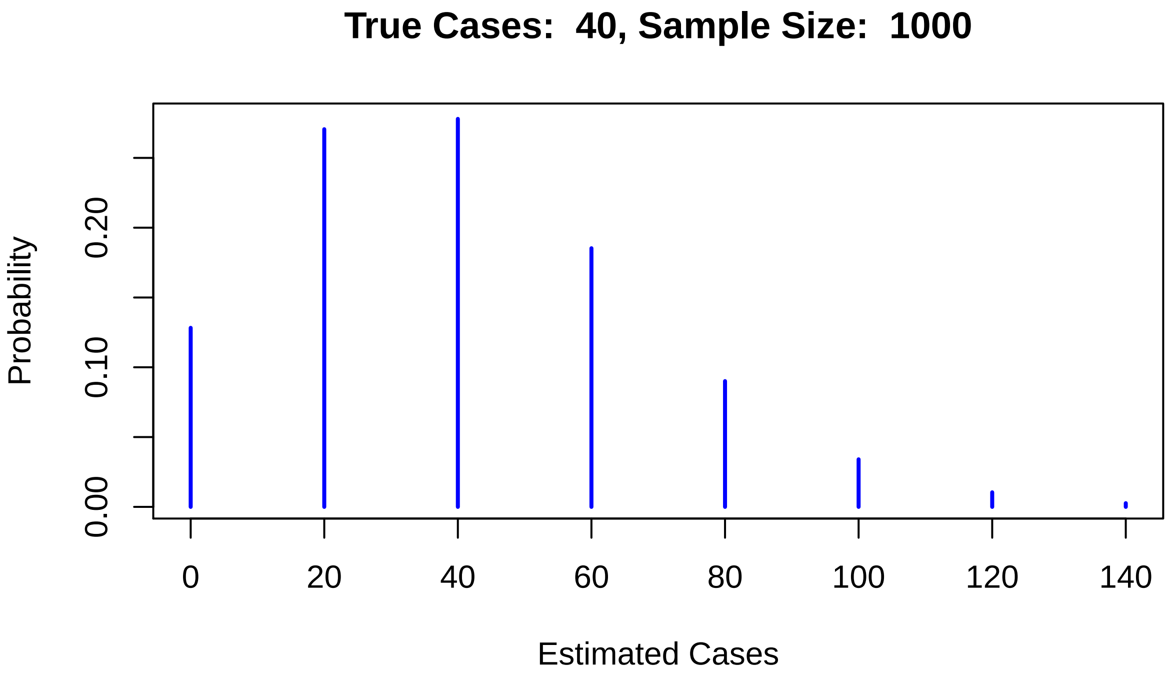
Because students are chosen at random, the estimated number of cases E may not equal the true number of cases C. Purely by chance, we might draw a sample of students in which no one is infected. In this case E would equal zero: an underestimate of C. We could also be unlucky in the opposite direction and, purely by chance, draw a sample that contains *everyone* at Oxford who is infected. In this case E would be an overestimate of C. So how can we determine whether E will provide useful information about C?

The answer is by considering *all possible samples* of students that we could draw. To make this more concrete, consider a simple analogy. A bowl (university) contains 20 marbles (students). Of these, 10 are red (positive for coronavirus) and 10 are blue (negative for coronavirus). It turns out that there are 15504 possible samples of 5 marbles that one could draw, each just as likely as any other. What fraction of these possible samples contain 5 red balls? It turns out that the answer is around 1.6%. While the details of the calculation are a bit different, the intuition is similar to that of flipping a fair coin five times: you are unlikely to get five heads in a row. In contrast, the chance of drawing 2 red and 3 blue marbles is high: around 34.8%. The numbers involved in this example are quite different from those in our real-life testing problem, but the basic intuition is the same: a random sample is likely to be similar to the population from which it was drawn.

Using this idea, we can calculate what is called the *sampling distribution* of our estimator E of the true number of cases C in the university.

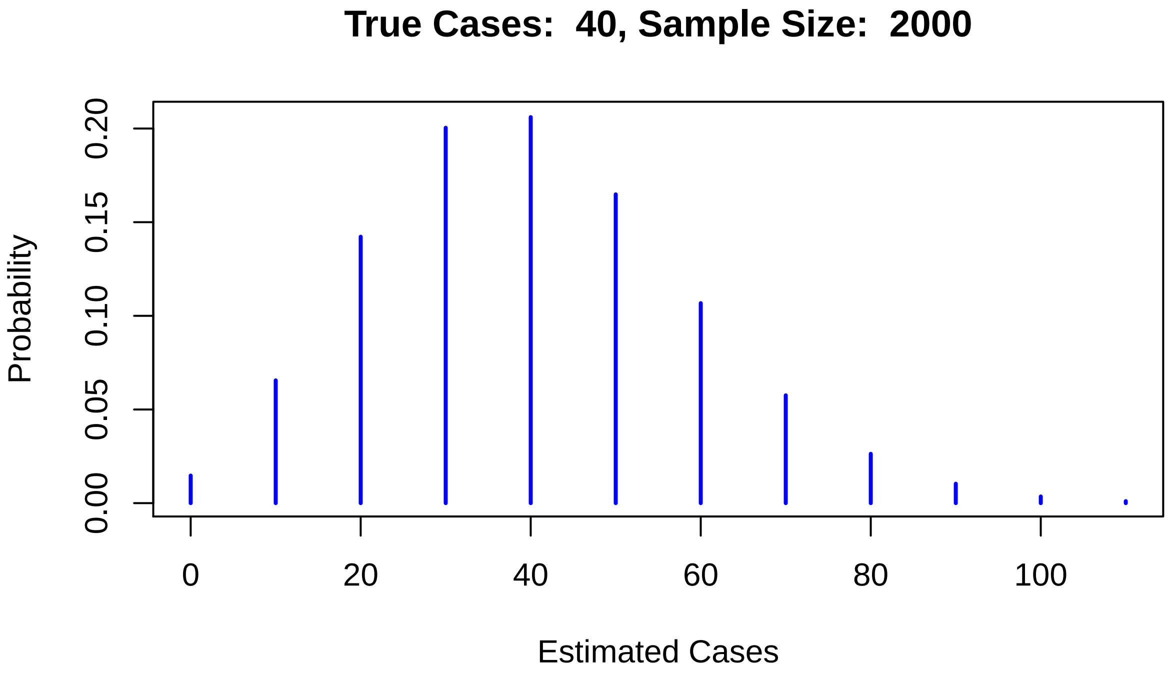
The two key advantages of random testing over other kinds of testing procedures, e.g. testing only symptomatic individuals, are as follows. First, neither over-estimates nor under-estimates predominate: the *average* value of $\widehat{C}$ is C. In statistical parlance we say that $\widehat{C}$ is an *unbiased estimator* of C. Second, by choosing an appropriate sample size n we can ensure that $\widehat{C}$ is very unlikely to be far from C.

This latter point is best understood using an example. Suppose that there are 40 coronavirus cases among 20,000 students. This amounts to a prevalence of 0.2%, which would be equivalent to two doublings of the ONS estimate of 0.05% from late August. Accounting for the fact that cases are currently more common among younger age groups, this may be a reasonable approximation to the state of play near the beginning of MT. The following figure shows the different values that the estimated number of cases $\widehat{C}$ could take on, along with their probabilities. In statistical parlance, this is called the "sampling distribution" of $\widehat{C}$. We can calculate these probabilities *exactly* because the randomness in $\widehat{C}$ is entirely under our control: it comes from the fact that students are randomly chosen to be tested.

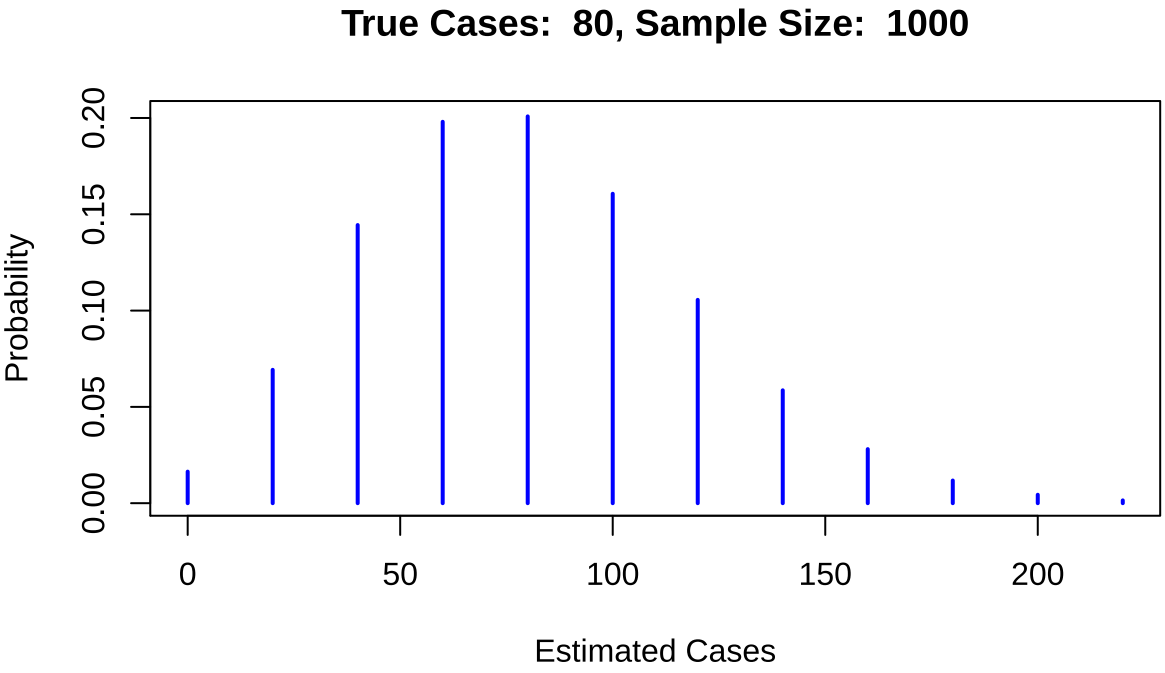


From the figure, we see that $\widehat{C}$ there is a 28% chance that our estimate $\widehat{C}$ will equal 40, the true number of cases. Moreover, there is a 73% chance that it will be between 20 and 60. There is, of course a chance that our sample will not contain any positives: $\widehat{C}$ equals zero with 13% probability.

For a given number of true cases in the population, increasing the sample size makes it more likely that $\widehat{C}$ will be close to C. For $n = 2000$ the sampling distribution of $\widehat{C}$ becomes more "bell-shaped." With this increased sample size, there is now a 82% chance that it will be between 20 and 60. The chance that we will fail to detect any cases falls to 1%.



For a fixed sample size, the probability of failing to detect any cases also falls as the number of true cases rises. Suppose, for example, that the number of cases in the population were to double from 40 to 80. Then, with a sample size of 1000, the probability that $\widehat{C} = 0$ would fall from 13% to 2%. As a rough guide, a sample size between one and two thousand tests per week should be sufficiently informative to detect cases when the prevalence is 0.2% or above.



A simple random sampling procedure is the easiest to explain and to use for back-of-the-envelope calculations, but there are several ways that it could be refined to increase the chance that $\widehat{C}$ will be close to C for a given sample size and true number of cases. The first would be to use a *stratified* rather than simple random sample. It is likely that coronavirus cases will tend to "cluster" within colleges, since the infection is most easily spread through frequent, close contact. If this is so, then a sampling procedure that allocates a *fixed* number of tests to each college (proportional to its size) and draws a simple random sample *within* each college will be more efficient than a simple random sample taken at the university level. Another possibility is to employ *adaptive sampling*, in which additional tests are allocated to colleges in which positive tests are detected in a "first wave" of tests. While the analysis of such a design is more complicated mathematically, it makes more efficient use of a limited number of tests if cases tend to cluster within colleges.

Pooled Testing

What about false positives and negatives?